Capítulo 5

Análisis de regresión

INTRODUCCIÓN

OBJETIVO DE LA REGRESIÓN

Determinar una función matemática sencilla que describa el comportamiento de una variable dados los valores de otra u otras variables.

DIAGRAMA DE DISPERSIÓN

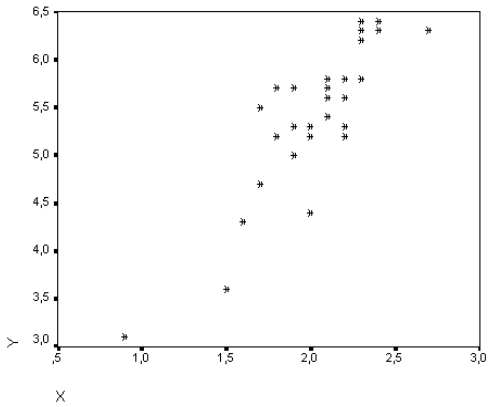

Figura1

Figura1: Diagrama de dispersión que relaciona la variable longitud (y) con una variable altura (x) de la concha Patelloida Pygmatea

Investigador

$\Downarrow$

Especificación de la forma funcional de la función de regresión

REGRESIÓN LINEAL SIMPLE

Suponemos un modelo en la forma

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \;\; ; \;\; i = 1, \ldots, n$$

- y_i : v.a. que representa la observación $i-$ésima de la variable respuesta, correspondiente al $i-$ésimo valor x_i de la variable predictiva X
- ε_i : Error aleatorio no observable asociado a y_i .

EJEMPLOS DE MODELOS DE REGRESIÓN SIMPLE

1) El consumo de gasolina de un vehículo, cuya variación puede ser explicada por la velocidad media del mismo. Podemos incluir en el término del error aleatorio el efecto del conductor, del tipo de carretera, las condiciones ambientales, etc.

2) El presupuesto de una universidad, cuya variación puede ser predicha por la variable explicativa número de alumnos. En el término del error aleatorio pueden incluirse el efecto del número de profesores, del número de laboratorios, de la superficie disponible de instalaciones, del número de personal de administración, etc.

ESTIMACIÓN POR MÍNIMOS CUADRADOS

$$\boxed{b_1 = \widehat{\beta}_1 = \frac{Cov(x,y)}{S_x^2}} \quad ; \quad \boxed{b_0 = \widehat{\beta}_0 = \bar{y} - b_1\bar{x}}$$

RECTA DE REGRESIÓN ESTIMADA

$$\boxed{\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i} \quad \text{o} \quad \boxed{\widehat{y}_i = \bar{y} + \widehat{\beta}_1(x_i - \bar{x})}$$

- $\widehat{\beta}_1$: la variación que se produce en $\widehat{y}$ por cada unidad de incremento en x

COEFICIENTE DE CORRELACIÓN LINEAL

Es una medida de la asociación lineal de las variables x e y

$$\boxed{r = \frac{Cov(x,y)}{S_x S_y}\ , \quad -1 \leq r \leq 1}$$

- Si $r = -1 \Rightarrow$ relación lineal negativa perfecta entre x e y
- Si $r = 1 \Rightarrow$ asociación lineal positiva perfecta entre x e y
- Si $r = 0 \Rightarrow$ no existe ninguna relación lineal entre x e y

ANÁLISIS DE LA VARIANZA

Si $\widehat{y}_i$ son estimadores de y_i

$$\Downarrow$$

$$y_i - \bar{y} = (y_i - \widehat{y}_i) + (\widehat{y}_i - \bar{y})$$

ECUACIÓN BÁSICA DEL NÁLISIS DE LA VARIANZA

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \widehat{y}_i)^2 + \sum (\widehat{y}_i - \bar{y})^2$$

$$SCT = SCE + SCR_{eg}$$

SCT : Suma de cuadrados total
SCE : Suma de cuadrados residual
SCR_{eg} : Suma de cuadrados de la regresión

Tabla ANOVA				
Fuentes de Variación	Sumas de Cuadrados	Grados de libertad	Cuadrados medios	F
Regresión	$SCR_{eg} = \sum (\widehat{y}_i - \bar{y})^2$	1	MCR_{eg}	$\frac{MCR_{eg}}{MCE}$
Error	$SCE = \sum (y_i - \widehat{y}_i)^2$	$n-2$	$MCE = \frac{SCE}{n-2}$	
Total	$SCT = \sum (y_i - \bar{y})^2$	$n-1$	$\frac{SCT}{n-1}$	

COEFICIENTE DE DETERMINACIÓN

Estadístico que representa la proporción de variación explicada por la regresión

Es una medida relativa del grado de asociación lineal entre x e y

$$R^2 = \frac{SCR_{eg}}{SCT} = 1 - \frac{SCE}{SCT} \;\; ; \;\; 0 \leq R^2 \leq 1$$

- Si $R^2 = 0 \Rightarrow SCR_{eg} = 0 \Rightarrow$ El modelo no explica nada de y a partir de x.
- Si $R^2 = 1 \Rightarrow SCR_{eg} = SCT \Rightarrow$ Ajuste perfecto: y depende funcionalmente de x .

★ Un valor de R^2 cercano a 0 $\Rightarrow$ Baja capacidad explicativa de la recta.

★ Un valor de R^2 próximo a 1 $\Rightarrow$ Alta capacidad explicativa de la recta.

EL CONTRASTE DE REGRESIÓN

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Fijado un nivel de significación α , se rechaza H_0 si $F_{exp} > F_{\alpha,1,n-2}$

EJEMPLO

La Patelloida Pygmatea es una lapa pegada a las rocas y conchas a lo largo de las costas protegidas en el área Indo-Pacífica. Se realiza un experimento para estudiar la influencia de la altura (x) de la Patelloida Pygmatea en su longitud (y) medidas ambas en milímetros. Se tienen los siguientes datos:

x	y	x	y	x	y	x	y
0.9	3.1	1.9	5.0	2.1	5.6	2.3	5.8
1.5	3.6	1.9	5.3	2.1	5.7	2.3	6.2
1.6	4.3	1.9	5.7	2.1	5.8	2.3	6.3
1.7	4.7	2.0	4.4	2.2	5.2	2.3	6.4
1.7	5.5	2.0	5.2	2.2	5.3	2.4	6.4
1.8	5.7	2.0	5.3	2.2	5.6	2.4	6.3
1.8	5.2	2.1	5.4	2.2	5.8	2.7	6.3

SOLUCIÓN

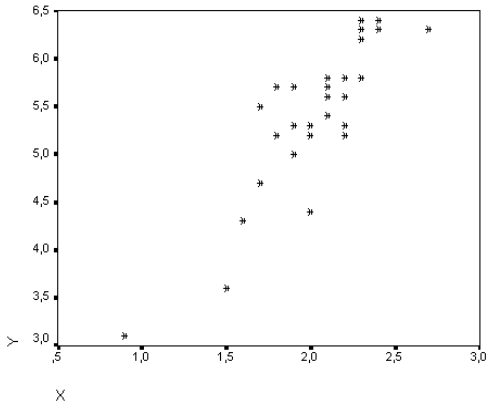

Figura1

Figura1: Diagrama de dispersión que relaciona la variable longitud (y) con una variable altura (x) de la concha Patelloida Pygmatea

Recta de regresión estimada

$$\widehat{y} = 1.36 + 1.99\ x$$

Coeficiente de correlación lineal

$$r = 0.8636$$

Coeficiente de determinación

$$r^2 = R^2 = 0.74$$

$$\Downarrow$$

El 74 % de la variabilidad de y puede atribuirse a una relación lineal con x

Contraste de regresión

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

A un nivel de significación del 5 %,

$$F_{exp} = 76.42 > F_{\alpha,1,n-2} = F_{0.05;1.26} = 4.23$$

Nótese además que el valor $p < \alpha$.

Rechazamos la hipótesis nula de no linealidad del modelo

REGRESIÓN LINEAL MÚLTIPLE

La v.a. y se relaciona con k variables explicativas $x_1, \ldots, x_k$

$$\Downarrow$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

■ Los parámetros $\beta_0, \beta_1, \ldots, \beta_k$ son estimados por mínimos cuadrados.

Para n observaciones podemos escribir:

$$\begin{array}{l} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_k x_{1k} + \varepsilon_1 \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_k x_{nk} + \varepsilon_n \end{array}$$

En notación matricial

$$Y = X\beta + \varepsilon$$

donde

$$X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ 1 & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \vdots & x_{nk} \end{pmatrix} \quad ; \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$$

y

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad ; \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

El vector de coeficientes β es estimado por mínimos cuadrados por:

$$\boxed{B = (X^t X)^{-1} X^t Y}$$

La ecuación ajustada de regresión resultante es:

$$\boxed{\widehat{Y} = XB}$$

ANÁLISIS DE LA VARIANZA

$$\boxed{y_i - \bar{y} = (y_i - \widehat{y}_i) + (\widehat{y}_i - \bar{y})}$$

ECUACIÓN BÁSICA DEL NÁLISIS DE LA VARIANZA

$$\boxed{\sum (y_i - \bar{y})^2 = \sum (y_i - \widehat{y}_i)^2 + \sum (\widehat{y}_i - \bar{y})^2}$$

$$\boxed{SCT = SCE + SCR_{eg}}$$

SCT : Suma de cuadrados total
SCE : Suma de cuadrados residual
SCR_{eg} : Suma de cuadrados de la regresión

Tabla ANOVA

Fuentes de Variación	Sumas de Cuadrados	Grados de libertad	Cuadrados Medios	F_{exp}
Regresión	$B^t X^t Y^t - \frac{1}{n}(\sum y_i)^2$	k	$CMR_{eg} = \frac{SCR_{eg}}{k}$	$\frac{CMR_{eg}}{CME}$
Error	$Y^t Y - B^t X^t Y$	$n-k-1$	$CME = \frac{SCE}{n-k-1}$	
Total	$Y^t Y - \frac{1}{n}(\sum y_i)^2$	$n-1$		

COEFICIENTE DE DETERMINACIÓN MÚLTIPLE

$$R^2 = \frac{SCR_{eg}}{SCT} = 1 - \frac{SCE}{SCT} \qquad ; \qquad 0 \leq R^2 \leq 1.$$

Representa la proporción de variación de y explicada por la regresión

- Si $R^2 = 0 \Rightarrow SCR_{eg} = 0 \quad \Rightarrow$ El modelo no explica nada de la variación de y a partir de su relación lineal con $x_1, \ldots, x_k$.
- Si $R^2 = 1 \Rightarrow SCR_{eg} = SCT \quad \Rightarrow$ Toda la variación de y es explicada por los términos presentes en el modelo.
- ★ Un valor de R^2 cercano a 1 $\Rightarrow$ Mayor cantidad de variación total es explicada por el modelo de regresión.

COEFICIENTE DE DETERMINACIÓN CORREGIDO

$$\overline{R}^2 = 1 - \frac{\dfrac{\sum e_i^2}{n-k-1}}{\dfrac{\sum (y_i - \overline{y})^2}{n-1}}$$

$$e_i = y_i - \widehat{y}_i$$

EL CONTRASTE DE REGRESIÓN

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \\ H_1 : \beta_j \neq 0 \quad \text{para algún} \quad j = 1, \ldots, k \end{cases}$$

Fijado un nivel de significación α , se rechaza H_0 si $F_{exp} > F_{\alpha,k,n-k-1}$

Bibliografía utilizada:

- ★ **Canavos, George C. (1988).** *"Probabilidad y Estadística. Aplicaciones y Métodos". Ed.: Mc Graw Hill.*
- ★ **Lara Porras A.M. (2002).** *"Estadística para Ciencias Biológicas y Ciencias Ambientales. Problemas y Exámenes Resueltos". Ed.: Proyecto Sur.*
- ★ **Milton, Susan (2002).** *"Estadística para Biología y Ciencias de la Salud". Ed.: Mc Graw-Hill.*
- ★ **Peña, Daniel (2002).** *Regresión y diseño de experimentos". Ed.:Alianza Editorial.*

- ♦ **Temporalización:** Dos horas